International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 14 (2015) pp 34063-34067

Chameleon Hierarchical Clustering Algorithm For Efficient Web Usage Mining

[1]Anupama Prasanthand[2]Dr. M. Hemalatha

[1]Research Scholar, [2]Professor
[12]Department of Computer Science, Karpagam University, Coimbatore

Abstract
The internet has progressed with a wide range of information over last few eras. This also brings the issue of information flooding and may cause the interruption to the user when accessing pertinent information. So it becomes a challenging task for the webmasters to provide the search options according to the individual user. Therefore, Web usage mining is an efficient technique where the helpful information is extracted from the user's history stored in the server logs and clusters those logs into a variety of labels. Hence the proposed system used the CHAMELEON clustering algorithm which processes the search logs of the user and cluster those logs into separate clusters based on the search category. Finally the chameleon algorithm is compared with other data clustering algorithms in terms of the Sum of Score Squared Error (SSE), rand index and F-measure respectively.

Keywords: information flooding, pertinent information, CHAMELEON clustering algorithm, web usage mining.

1. Introduction

In recent years, internet world has become tremendously popular and its development is very rapid. The huge amount of available information on the internet is utilized by the people for their various intend. While searching the specific information on the web, it is significant for the user to get the information in less time [1]. Thus, Web mining is a kind of data mining approach to automatically extract information and determine the information through the analysis of Web structure, Web usages, and Web contents. Web usage mining is one among the web mining techniques that determine the web usage data about the usage patterns to recognize and oblige the requirement of the web based applications [2]. Predicting the users' browsing history is one of the web usage mining techniques. Usage information captures a source of web users along with their browsing performance on a web site.

Web Usage Mining is the method of extracting useful and helpful patterns from the web log files. There are various methods namely clustering, SVM, Markov model, Page ranking, Association rule mining, Modified Markov model, Markov model with clustering, Association rule mining etc. has been utilized for web page prediction [3]. Clustering is the method of combining the specified set of unlabeled patterns into clusters so that the similar patterns are consigned to one cluster. Each pattern is signified by a vector of many factors [4]. The traditional clustering algorithm uses closeness or interconnectivity to combine the clusters of similar patterns. Chameleon algorithm deliberates both closeness and interconnectivity for merging the related cluster pair of similar patterns. Chameleon also uses a methodology to ideal the interconnectivity and a closeness degree between the cluster pairs.

Two clusters are combined only if the closeness and inter-connectivity between tthem are high relative to the internal interconnectivity of the closeness and cluster items within the clusters. The Chameleon is a clustering algorithm based on the K-nearest neighbor (KNN) Algorithm. The basic Chameleon Algorithm process is shown in figure 1.

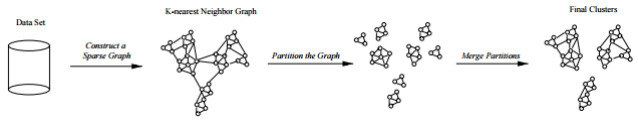

Figure 1. Chameleon Algorithm

Chameleon algorithm first creates a KNN graph of the given data set and uses a multilevel graph partitioning algorithms such as hMETIS to find the initial sub-clusters then determine the relative interconnectivity between relative closeness and two clusters within a cluster. This cluster merging process continue until achieving the specified number of clusters created. Chameleon uses a two level of method to find the final clusters. In initial level, it utilizes a graph partitioning algorithm and in second level cluster the data items into numerous relatively tiny sub-cluster.This paper, presents Chameleon algorithm for web data clustering in terms of web usage mining. This process is applicable to any type of data and also similarity matrix can be created. The Chameleon algorithm comprises clusters of various densities, shapes, noise, size and artifacts and it also contain points in 2D space.

2. Related work

Web usage mining is one of the computational intelligence application which haveused as a tool for creating adaptive web sites, web site management, business, personalization, network traffic flow analysis and support services, and so on, in web based applications. Thus, in [5] author introduced a process for finding user access patterns using user's behavior and session. This process is combined with three basic operations of web usage mining operation such as preprocessing, pattern discovery and pattern extraction.

Finally, combined association rule mining and cluster effort is employed for pattern discovery.

One of the data mining application is web usage mining is used to web log data repositories. Finding the user access pattern from the web access log. Thus, in [6] and [7] author create a novel approach to finding the user's order of occurrence of the visits and behavior of the user page visits by using novel rough set Dbscan clustering algorithm. Web data clusters are created by using the rough set Similarity Upper Approximation (SUA). The experimental evaluation of rough set Dbscan clustering approach show promising results to compare with rough set agglomerative clustering by using MSNBC web navigation dataset.

In [8] author presents a cluster approach for web usage mining, which initialize the cluster center by using information entropy and weighting parameters are used to adjust the cluster center location. Web data clustering is used to find the groups which share common behavior and interest by analyzing the data acquired from the web servers. To enhance the clusters by using Fuzzy C Mean (FCM) algorithm,it is used to find the user access patterns from the web access log and MSNBC web navigation is used for forming web data clusters.

Cluster analysis is a primary approach for conventional data analysis and numerous clustering approaches have been recognized which needs a number of clusters to be specific in advance and is dependent on initial starting points. Thus, in [9] author present a novel algorithm to discover data cluster for nominal and numerical data. The Apriori algorithm creates a huge number of candidate set that is not efficient for both data.

Web usage mining is the sort of web mining approach that includes the automatic discovery of user access patterns from more web server. Thus, in [10] author analyzes the pattern, utilizing various algorithms, for example Hash tree, Apriori, Fuzzy and finally utilized enhanced Apriori algorithm to give the outcome for Crisp Boundary problem with higher enhanced efficiency while comparing with other existing algorithms.

3. The Usage Mining System uses Chameleon algorithm

The main aim of the proposed work is to find web user clusters from the web server log files. This web usage mining approach is partitioned into two phases. The first phase is splitted into user identification stage and session constructions where both the stages are considered as data preprocessing stage. In the second phase, Chameleon cluster algorithm is utilized to create the cluster.Theweb usage mining system by utilizing the Chameleon algorithm is shown in the Figure 2.

The web log data from the web server is normally voluminous and diverse. This raw log data must be collected in an integrated, comprehensive and consistent form, so as to be utilized for pattern discovery [11][12]. Like the other variousdatamining application, the users are uniquely identified and the collected log data of the separate users are preprocessed. This preprocessing stage comprises of filtering, removing the irrelevant and redundant data; and makeover any inconsistencies and finally the user sessions are created. In order to identify the individual users, the existing approaches consumes various methods. The individual users are recognized by the combination of browser and the IP address of the particular user [13] [14].After this process the sessions are created from the individual user's data. Error records, requests for the multimedia and image files are removed from the web server log files by using a tool called web log filter. Then the timestamp, IP address, user agent, referrer and request are recollected for the further processing.

After data preprocessing, employ the subsequent conditions. The problem statement defined as $L = \{i_1, i_2, \ldots, i_m\}$ be a set of items then consider D be a set of transactions and each transaction item T is subset of L. For each webtransaction$T$, X set of items are comprised in L. As if different users recurrently access the same sequence of web pages, a corresponding sequence of log entries will display in the log files, and this sequence can notidentify an access pattern.

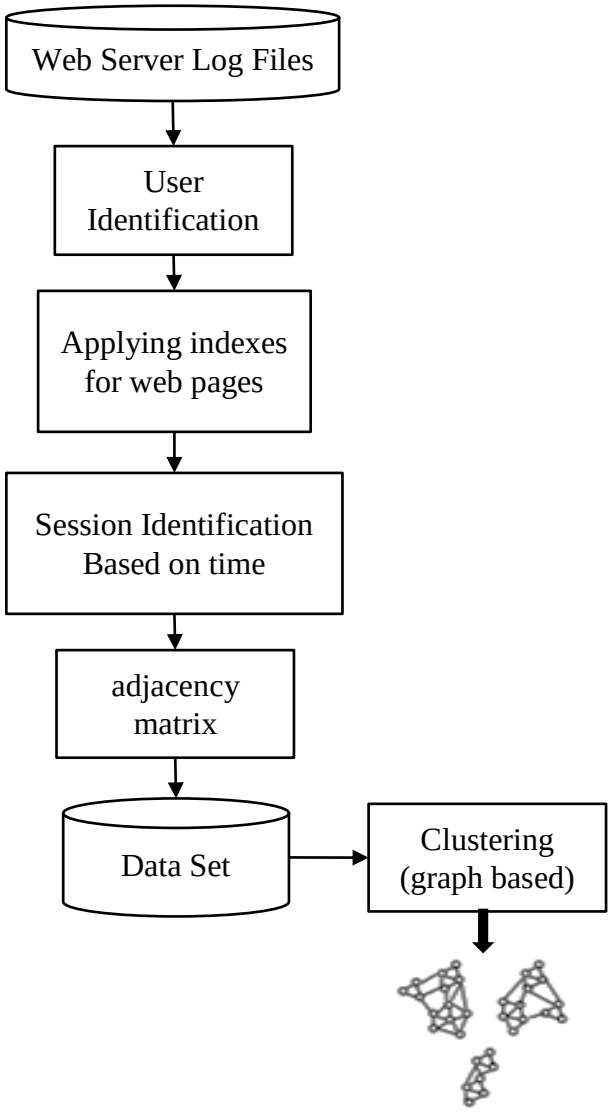

Figure 2 Web Usage Mining System using Chameleon algorithm

The log was gathered from 00:00:00 April 1, 2015 through 23:60:60 May1, 2015, a total of 31 days. The information extraction is done for instances most frequently accessed and requested files are extracted by analyzing the log files. The table 1 shows some indexing web pages.

Table 1 web page index

INDEX ID	FILE NAMES
1	http://www.science.gov/
2	http://www.science.gov/browse/w_121G1.htm
3	http://www.science.gov/browse/w_113N.htm
4	http://m.science.gov/scigovmobile/
5	http://www.science.gov/helpNew.html
6	http://www.science.gov/helpNew.html#advanced
7	http://www.science.gov/helpNew.html#Search Tips

In session construction stage, the Time based method is utilized for identifying the user sessions.A session is created based on a time period. Each and every unique session is also given a unique index file, that contains the web pages refered in that specific session and session number. The Threshold timeout session is set as 20 minutes. The session construction algorithm as shown in table 2.

Table 2 Session Construction

Session construction

Input: indexed file and User identified log file
Output: Session files
Step: 1 For each User U based on Distinct $(ip + browser)$
Step: 2 For each request r of User U
Step 3: If $(time\_of\_current\_req - time\_of\_first\_req < 20\ mins)$
Add this user request to the current session
Else
Step 4: Create a new session and Write session details $(ip, time, req)$ to session file
$curr\_req\_time = curr\_req\_time - 20$
End For
End For

The adjacency matrix is created by using session and requested pages. This matrix is formed by considering those pages as column and session as row. The data point in this adjacency matrix is given as input to the next clustering process. The clustering operation is done by using chameleon algorithm.

Chameleon utilizes KNN graph method to characterize its objects (datapoints or search items) in which it defines more natural clusters. The neighborhood is signified intensely in a dense region, whereas it is characterized more extensively in a spare region. Partitioning and merging are the two phases of the Chameleon algorithm. Duringthe partition phase, the cluster graph makes use of multi-level graph partitioning algorithm calledhMETIS which is software package of circuit design and used for large hypergraph partitioning. In hypergraph, once create cluster into k partition, the bad clusters are automatically eliminated by using cluster fitness value, is as follows

$$fitness\ (C) = \frac{\sum_{e \subseteq C} Weight\ (e)}{\sum_{|e \cap C|>0} Weight\ (e)} \quad (1)$$

Where e be a set of vertices signifying a hyperedge and C be a set of vertices signifying a partition.

Once the good partitions are found, is examined to filter out vertices which are not correctly connected to the rest of the vertices of the cluster partition. The cluster connectivity function of vertex v in C is defined as follows

$$Connectivity = \frac{|\{e | e \subseteq C,\ \ v \in e\}|}{e | e \subseteq C} \quad (2)$$

High connectivity value recommends that the vertex has numerous edges connecting good quantity of the vertices in the partition.The hMETIS improves the efficiency and speed of chameleon cluster algorithm.

Chameleon defines i, j lines the similarity between each pair of clusters C_i and C_j to their relative closeness $RC\ (C_i C_j)$ and relative inter-connectivity $RI\ (C_i C_j)$define as absolute inter-connectivity between two clusters. The relative inter-connectivity equation is as follows

$$RI\ (C_i C_j) = \frac{\left|EC_{\{C_i C_j\}}\right|}{\frac{\left|EC_{C_i}+EC_{C_j}\right|}{2}} \quad (3)$$

Where Edge-Cut (EC), is described as a sum of weight edge, the edge that connect $C_i$to C_j. The relative closeness equation is as follows

$$RC\ (C_i C_j) = \frac{\bar{S}EC(C_i, C_j)}{\frac{|C_i|}{|C_i|+|C_j|}\bar{S}EC(C_i) + \frac{|C_i|}{|C_i|+|C_j|}\bar{S}EC(C_j)} \quad (4)$$

Where $\bar{S}EC(C_i)$ and $\bar{S}EC(C_j)$ are the average weights of the edges that belong in clusters and connectivity vertices C_i and C_j and number of data points in each cluster are defined as $|C_i|, |C_j|$.

Chameleon algorithm can overcome the limitation of the existing clustering algorithm in terms of relative closeness whereas the existing clustering algorithm only focused on the absolute closeness. In the merging phase of the chameleon algorithm, the clusters are merges with each other by using the relative closeness and relative inter-connectivity. Sothat the ensuing cluster results in a constant amount of closeness among the items in the cluster.

4. Result and discussion

To evaluate the performance of Chameleon algorithm based clustering in a very controlled manner, multidimensional data

sets were utilized. To measure the clustering accuracy of the Chameleon algorithm, the metrics such as Sum of Squared Error (SSE) and Rand Index are taken into account.

(A) Metric

The Rand index is a measure utilized to compare the clustering structure of $C_i$and C_j. The Rand index equation is defined as follows

$$Rand\ index = \frac{w + x}{w + x + y + z} \quad (5)$$

Where w denotes the number of instances in the both the clusters, z denotes the number of instances in different cluster, y denotes the number of instances assingned in the cluster C_j but not in C_i and finally x denotes the number of instances assingned in the cluster $C_i$but not in C_j.

Sum of Squared Error (SSE) is the most widely utilized and simplest criterion measure for clustering. Cluster compactness is measured by using SSE. It is calculated as

$$SSE = \sum_{K=1}^{K} \sum_{\forall x_i \in C_k} \|x_i - \mu_k\|^2 \quad (6)$$

Where μ_k is the vector mean of the cluster k, C_k is defined as a set of cluster k instances.The μ_k is calculated by using below equation

$$\mu_{k,j} = \frac{1}{N_k} \sum_{\forall x_i \in C_k} x_{i,j} \quad (7)$$

Where,$N_k = |C_k|$is the number of instances corresponding to the cluster k.

The F-measure is calculated by merging recall and precision values of clusters.

$$precision\ (i,j) = \frac{c_{ij}}{c_j}$$

$$recall(i,j) = \frac{c_{ij}}{c_i}$$

The f-measure, $F(i,j)$ of a class i concerning cluster j is then defined as

$$F(i,j) = \frac{2 * precision\ (i,j) * recall(i,j)}{precision\ (i,j) + recall(i,j)} \quad (8)$$

(B) Results

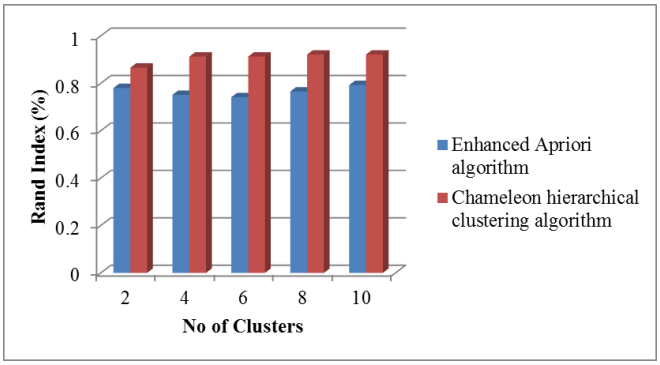

Figure 3. Rand Index

Figure 3 shows the comparison result of two algorithms such as Enhanced Apriori algorithm and proposed Chameleon hierarchical clustering algorithm in terms of Rand Index measures. If the Rand Index value is nearly equal to one, then the clustering accuracy is high. The proposed method shown promising results while comparing with the existing Enhanced Apriori algorithm.

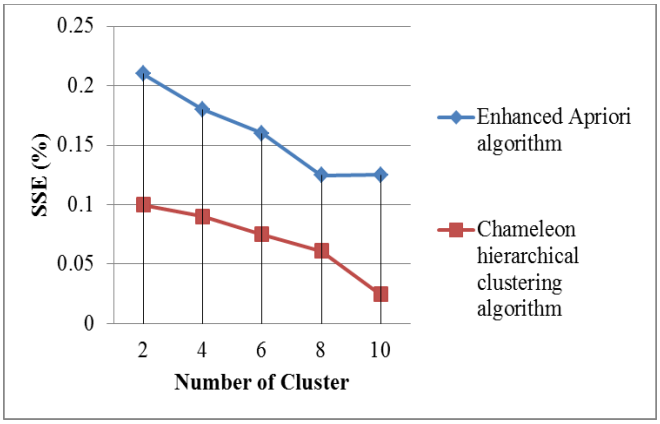

Figure 4.Sum of Squared Error (SSE)

The results of SSE between clustering methods such as Enhanced Apriori algorithm and proposed Chameleon hierarchical clustering algorithm shown in Figure 4 in terms of a user session matrix by using SSE measures. It shows that the proposed method has less error value when compared to Enhanced Apriori algorithm methods.

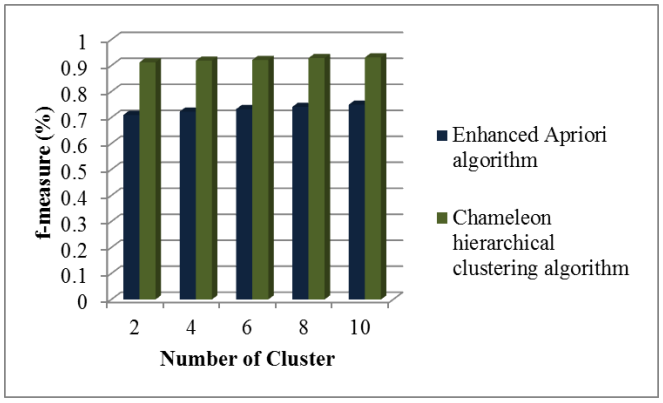

Figure 5. F-Measure

The F-measure value is high when there is correct cluster formation. Thus, Proposed Chameleon hierarchical clustering algorithm have a high clustering accuracy than existing Enhanced Apriori algorithm clustering methods. The F-measure performance result is shown in Figure 5 and depicts that the proposed method shows the promising result when compared with the existing algorithm.

5. Conclusion

The process of gathering the similar data in a group is known as clustering. Many algorithms exist for the clustering, even the data used in various applications, but still experiencing many drawbacks. So in order to conquer these drawbacks, the Chameleon hierarchical clustering algorithm is used. Identifying the closeness between the data points in the clusters and inter-connectivity between any two clusters are used to find the similar cluster and this is the most important behavior for the chameleon algorithm. The time oriented approach is utilized initially to generate the session. Then these generated sessions are converted into the data points and the adjacent matrix. These data points use the MATLAB to be plotted on the plane and the chameleon clustering algorithm identifies the clustering among the data points.

References

[1] Richa Patel, Akshay Kansara, "Web Usage Mining: A Survey on User's Navigation Pattern from Web Logs", International Journal for Scientific Research & Development (IJSRD), Vol. 2, Issue 09, 2014.

[2] P. Sampath, Prabhavathy M. "Web Page Access Prediction Using Fuzzy Clustering By Local Approximation Memberships (FLAME) Algorithm",ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 7, 2015.

[3] Karuna Katariya,Rajanikanth Aluvalu, "Agglomerative Clustering in Web Usage Mining: A Survey"International Journal of Computer Applications,Vol.89, No 8, 2014.

[4] Rupinder Kaur, Simarjeet Kaur, "A Review: Techniques for Clustering of Web Usage Mining",International Journal of Science and Research (IJSR),Vol.3 Issue 5, 2014.

[5] Shaily G.Langhnoja, Mehul P. Barot, Darshak B. Mehta, "Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery",International Journal of Data Mining Techniques and Applications,Vol. 02, Issue 01, 2013.

[6] K. Santhisree, A. Damodaram, SV Appaji, "An Enhanced Dbscan Algorithm to Cluster Web usage Data using Rough Sets and Upper Approximations",International Journal of Computer Science & Communication, Vol. 1, No. 1, pp. 263-265, 2010.

[7] K.Santhisree, A.Damodaram, "Clustering on Web usage data using Approximations and Set Similarities", International Journal of Computer Applications, Vol. 1 – No. 4, 2010.

[8] K.Suresh, R.Madana Mohana, A.RamaMohanReddy, "Improved FCM algorithm for Clustering on Web Usage Mining",International Journal of Computer Science (JCSI), Issues, Vol. 8, Issue 1, 2011.

[9] Pooja Sharma, Rupali Bhartiya, "An Efficient Algorithm for Improved Web Usage Mining",International Journal of Computer Technology & Applications,Vol 3 (2), 766-769, 2012.

[10] S.Veeramalai, N.Jaisankar, A.Kannan, "Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy ",International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, 2010.

[11] Dushyantsinh B. Rathod, Mr. Ramesh Prajapati, DrSamratKhanna, "Emerging Trends of Clustering Techniques In Web Usage Mining",IJSART, vol.1, Issue 4, 2015.

[12] L.K. Joshila Grace, V.Maheswari, DhinaharanNagamalai, "Analysis Of Web Logs And Web User In Web Mining " International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, 2011.

[13] G.S Vinothkumar, J.jamet, N.Kamal, "Design and Implementation of a Novel Webpage Ranking Algorithm for improved Web Search",International Journal of Inventions in Computer Science and Engineering,Vol.1, Issue 3, 2014.

[14] R.Pratheeba, R. Purushothaman, "Modeling Smarty Web Search Engine Using Xml Clustering",International Journal of Inventions in Computer Science and Engineering,Vol.1 Issue 2 2014